Using Machine Learning Techniques to Increase the Power of TSLS: On the Effect of Police on Crime

Shopnavo Biswas, Tanmay Gupta, David Tang

Abstract

Since the publishing of Levitt's seminal 1995 crime paper, wherein he instrumented for the causal effect of police on crime using mayoral/gubernatorial elections, there have been several criticisms of his paper published. One particular criticism is from McCrary (2002), wherein McCrary uses updated data to demonstrate that while the instrument is correlated with policing, the second stage is too weak to yield statistically significant results. In this paper, we investigate modifications to the standard TSLS method of providing estimates using an instrumental variables approach (IV). In particular, we investigate Bayesian Additive Regression Trees (BART-IV) and Random Forest (RF-IV) methods. We find that both BART-IV and RF-IV provide tighter estimates of the true parameter than TSLS, but the Bayesian inference in BART-IV is not well calibrated to a frequentist setting. We use RF-IV to instrument for the effect of policing on all different categories of crime, and find no statistically significant relationship at the 95% level. We ultimately construct a more powerful reproduction of McCrary's criticism of Levitt's paper, arguing that when simultaneity is factored in, policing does not seem to significantly alter crime rates.

1 Introduction

Levitt (1995) attempts to answer the much-debated question of whether increased policing leads to lower crime rates. OLS regressions suffer from a simultaneity problem in this application – crime rates and policing both have a causal effect on each other and mere correlations do not reflect the direction of the causality. Additionally, there is also an omitted variables problem, since unobservable facts about geography and culture might affect both policing and crime rates. In order to overcome these problems, Levitt uses an instrumental variables (IV) strategy, where he exploits the fact that electoral cycles in US cities exogenously affect police hiring. In particular, using a panel with crime-related data from several US cities between the years of 1970 and 199, Levitt uses indicators for whether the particular year was a mayoral or gubernatorial election year as instruments for police hiring. Levitt's two-stage least squares (TSLS) estimates for the effect of police on violent and property-related crime, as presented in his paper, are negative and statistically different from zero at a 95 percent confidence level. McCrary (2002) discovered mistakes in Levitt's original analysis, and found that the actual TSLS estimates were much smaller and magnitude and not statistically different from zero. McCrary's paper concludes that Levitt's instruments are weak – while they significantly predict policing, they are insufficiently predictive of crime rates to lead to causal identification in Levitt's data.

There has recently been a large literature on using machine learning (ML) methods in instrumental variables, particularly in applications with weak instruments. Bai and Ng (2010) present an approach in which estimated factors can be used as valid instruments for endogenous regressors. Chen et al. (2020) use a technique involving an ML prediction method (e.g. random forest or neural network) and a sample-splitting procedure to reduce standard errors in the IV strategy used by Ash et al. (2018). On the other hand, Angrist and Frandsen (2019) attempt to use random forests to predict the first-stage of the TSLS approach of Angrist and Krueger (1992), but find results to be no different to those obtained through regular TSLS. In general, as noted by Chen et al., machine learning approaches can be particularly useful when instruments are feared to be weak in a linear first-stage, since they can fit a far more general first-stage than linear regression allows for.

To our knowledge, ML methods have so far not been applied to Levitt's data on police and crime. In this paper, we examine whether ML methods can improve Levitt's analysis. We focus on two methods. Our first attempt is to use a random forest to predict the first-stage, and then use these fitted values in a linear regression in the second stage. Our second attempt is to use a Bayesian ensemble-of-trees method, Bayesian Additive Regression Trees (BART), to investigate a non-parametric estimation of both the first and second stages of IV. For both attempts, we run Monte Carlo simulations to assess the performance of these methods, and then apply them to the actual crime data. While both methods successfully produce coefficient estimates with smaller standard errors than pure IV, our simulations reveal that BART grossly underestimates standard errors, which renders its performance on the actual dataset invalid. Random forest, on the other hand, provides more reasonable standard errors, so we accept the final results we obtain by using it on the real data.

The remainder of the paper is organized as follows. In section 2, we discuss our data, our regression specifications, and more details on the ML methods. In section 3, we present and discuss our Monte Carlo simulation results. In section 4 we present our empirical results, and in section 5 we conclude.

2 Data and Methodology

We use replication data provided by McCrary (2002), which contains crime and police data for 59 US cities between the years of 1970 and 1999. The dataset also contains the number of police officers sworn in to the police force of each city each year, and year indicators for mayoral and gubernatorial election years. There is data on the number of crimes committed for seven crime categories: murder, rape, assault, robbery, burglary, larceny, and auto (car) theft. As in Levitt (1995), we group these crimes into two categories: violent crimes – murder, rape, robbery, and assault – and property crimes – burglary, larceny, and auto theft. We note that the indicator for mayoral elections in McCrary's data is not exactly the same as in Levitt's original data. As mentioned in his paper, McCrary suspected that there were errors in Levitt's mayoral-election year indicators (years wrongly counted as election years, election years omitted) and independently created corrected more accurate indicators.

In its most general form, we are interested in estimating the model

$$\Delta \ln(C_{it}) = \beta(\Delta \ln(P_{it})) + h(\mathbf{W}_{it}) + \varepsilon_{it}$$

where i subscripts city and t subscripts year. In this equation, C denotes the crime rate per $100,000$, P denotes the number of sworn officers per $100,000$, $\mathbf{W}$ is the same vector of observed confounders used in Levitt (1995), and h is some (possibly non-linear) function. Our object of interest is β, the causal effect of police on crime. We take the log differenced variables so that we are estimating elasticity.

Due to selection on unobserved confounders, this is not a conditional expectation model, i.e., $\mathbb{E}[\varepsilon_{it}|P_{it}, \mathbf{W}_{it}] \neq 0$. This means running a simple regression will not return the causal parameter of interest. To address this, we use instrumental variables with mayoral and gubernatorial elections as our instruments. In the instrument variables approach, we start by estimating the first stage

$$\Delta \ln(P_{it}) = h(Z_{it}, \mathbf{W}_{it}) + \varepsilon_{it}^{(1)}$$

where Z is our instruments. By exogeneity assumptions that make Z a valid instrument, this is a conditional expectation model. We can then use the fitted values from the first stage to estimate the causal parameter in the original model. Intuitively, this two stage estimation uses exogenous variation in Z as a proxy for variation in P, giving us a conditional expectation model that we can estimate with some regression framework.

In the classical TSLS method, we assume that f and h are both linear so that both the first and second stage reduce to an ordinary least squares regression. In RF-IV, we estimate $h(Z_{it}, \mathbf{W}_{it})$ as a nonparametric model with an ensemble of decision trees. We keep the assumption of linearity in the second stage. Finally, in BART-IV, we adopt a Bayesian approach and estimate both f and h using ensembles of nonparametric decision trees.

We need to take special care when obtaining standard errors for our estimates. Specifically, we need to consider the uncertainty with respect to our first stage estimates as well as uncertainty with respect to our second stage estimates. In general,

this means that the IV estimates will have a larger standard error than the OLS estimates.

For TSLS, we have an analytical form for the standard errors. To our knowledge, there is no analytical form for the standard errors of RF-IV. To obtain standard error estimates, we resort to bootstrapping. More specifically, we compute the estimate several times on data resampled with replacement. We then take the standard error of the estimate as the empirical standard deviation of the estimate in the bootstrap samples. The idea is that resampling the data with replacement approximates sampling from the true distribution.

BART avoids this issue of standard errors by adopting a Bayesian approach to the problem. To do so, we need to specify priors on all of our parameters. Specifically, we need to specify priors on $f, h, \beta, \varepsilon^{(1)}$, and $\varepsilon^{(2)}$ where $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ denote the error terms in the first stage and the second stage respectively.

We place a bivariate normal prior on $\varepsilon_{it}^{(1)}$ and $\varepsilon_{it}^{(2)}$ parameterized by mean μ_{it} and covariance Σ_{it}. To limit the flexibility of the model, we restrict the set of mean and covariance parameters in the bivariate normal priors to a finite set by modeling (μ_{it}, Σ_{it}) as draws from a discrete distribution governed by a Dirichlet mixture process. We place agnostic normal priors on f and h and model their likelihood with the BART methodology developed by Chipman et al. (2010). Finally, we place a normal prior on β. This approach follows the model in McCulloch et al. (2021).

After these priors have been specified, we can obtain empirical samples from the posterior using a Gibbs sampler. In Gibbs sampling, we take successive draws from the full conditionals to obtain realizations from a Markov chain whose stationary distribution is the posterior. These draws from the posterior can then be used to approximate the posterior mean and standard deviation.

We note that the TSLS model differs slightly from the one presented in Levitt (1995), where a more restrictive specification (with sharing of parameters across crime categories) is used. Our focus here is not replication, but on increasing the power of the model i.e. reducing the standard errors in our coefficients for a tighter estimate.

3 Simulation Results

To assess the performance of the methods of interest against OLS and two-stage least squares, we run Monte Carlo simulations with the following data generating process (DGP). In this DGP we have one binary instrument and one continuous treatment. We also generate three “confounders", variables which are correlated with both the outcome and treatment but are left out of the regressions to simulate an omitted variables situation, which we expect to be the case in Levitt's data. The treatment is a linear combination of the confounders and treatment with normal noise. When calculating the outcome variable Y, the coefficient on the treatment is three times larger than the coefficient on the confounders, all of which are equal. We also add normal noise to the outcome Y.

Let N be the number of datapoints desired. Let $\mathbf{X}$ be a $N \times 4$ matrix with $\mathbf{X}_0$ the treatment and $\mathbf{X}_i,\ 1 \leq i \leq 3$ the confounders each an $N \times 1$ vector. Let $\mathbf{Z}$ be an $N \times 1$ vector of the instrument. Then we use the following steps:

1. Generate $\mathbf{X}_i,\ 1 \leq i \leq 3$ independently from a standard normal distribution.
2. Generate $\mathbf{Z}$ from N draws of a bernoulli distribution.
3. Let $\mathbf{A}$ be a 3 x 1 draw of independent standard normal variables.
4. Then $\mathbf{X}_0 = \mathbf{N}(5Z, 1) + \mathbf{X}_{1,2,3}\mathbf{A} + \mathbf{U}$, where $\mathbf{N}(Z, 1) = [a_0, a_1, ...a_N]$ such that each $a_i \sim N(Z_i, 1)$ and $\mathbf{U}$ is a $N \times 1$ vector of standard normal error terms. Note that the notation $\mathbf{X}_{1,2,3}$ denotes a matrix with columns X_1, X_2, and X_3.
5. Initialize $\beta = [3, 1, 1, 1]$
6. Then $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, where $\mathbf{U}$ is standard normal noise [1]

3.1 Random Forest (RF-IV)

We use the DGP described in Section 2 to draw samples of X and Y 100 times. For each sample, we store the estimate for the parameter, as well as the 95% confidence interval (CI). We calculate the coverage as the proportion of samples in which the true parameter is contained inside the CI. We calculate the SD as the standard deviation

[1] Not necessarily the same as the $\mathbf{U}$ in Step 4, we just use the same notation again

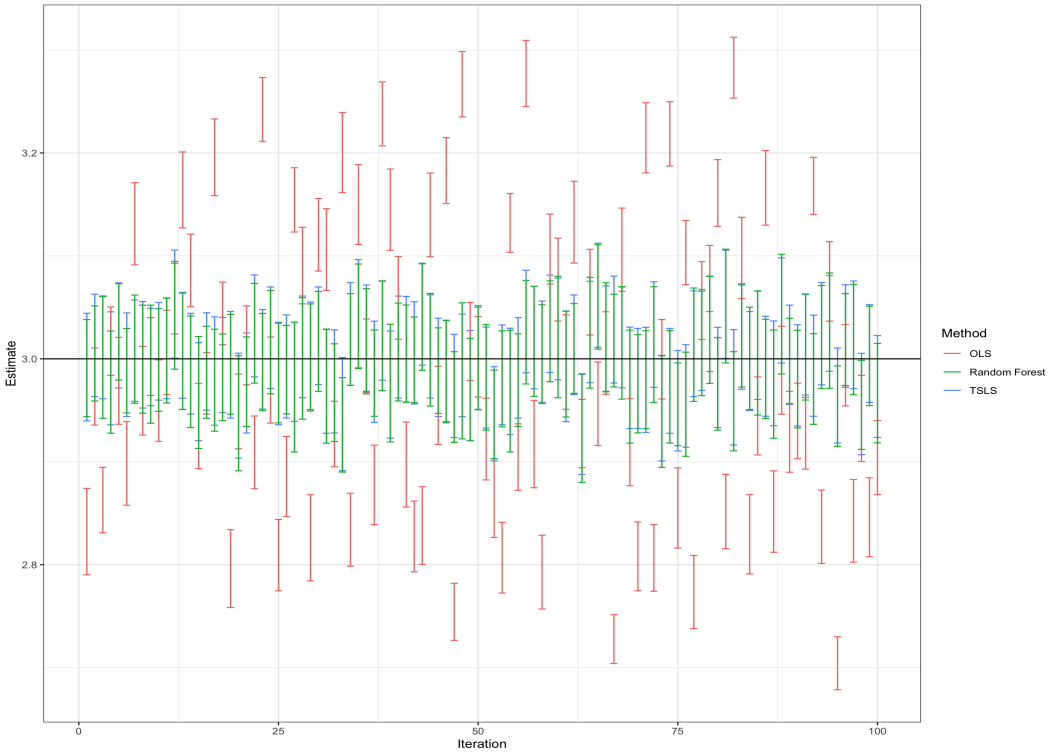
of the parameter estimates over all 100 samples.

Table 1: Monte Carlo Results for RF-IV

Method	Mean Estimate	SD	Coverage
OLS	2.991	0.139	0.17
TSLS	3.001	0.025	0.95
RF-IV	2.996	0.026	0.93

The table above reports Monte Carlo results for OLS, TSLS, and Random Forest IV using the DGP, for a total of 100 draws. 'Mean Estimate' refers to the sampling mean of estimated coefficient on the regressor of interest, 'SD' is the sampling standard deviation of the mean. For each method in each draw, we calculated 95% normal confidence intervals, and 'Coverage' reports the proportion of those which contained the true value, in this case calibrated to 3.

Figure 1: RF-IV Monte Carlo parameter estimates with confidence intervals across samples



We observe low coverage in OLS because of the simultaneity issue previously described. By contrast, we observe high coverage across the samples for both TSLS and RF-IV, and we see lower variance in the parameter estimates, suggesting that these estimates are less biased. This is heuristically confirmed by the fact that the mean estimate for TSLS and RF-IV are both within 0.005 of the true parameter value while OLS is within 0.01.

3.2 Bayesian Approach (BART-IV)

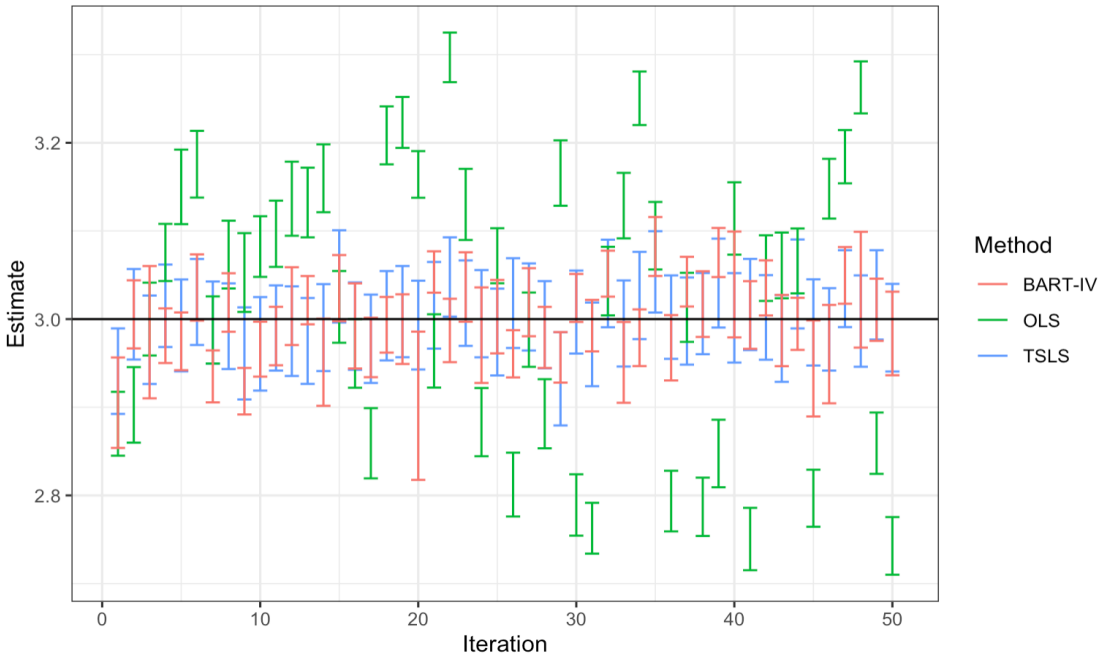
Next, we use Bayesian Additive Regression Tress (BART-IV), as described in section 2, on the same DGP for a total of 50 Monte Carlo draws. We used the ivbart implementation provided by McCulloch et al. (2021) with its default specifications for the priors.

Table 2: Monte Carlo Results for BART-IV with naïve DGP

Method	Mean Estimate	Average SE	SD	Coverage
OLS	3.022	0.019	0.152	0.12
TSLS	3.002	0.025	0.025	0.92
BART-IV	2.995	0.020	0.041	0.68

The table above reports Monte Carlo results for OLS, TSLS, and BART IV using this DGP, for a total of 100 draws. 'Mean Estimate' refers to the sampling mean of estimated coefficient on the regressor of interest, 'Average SE' is the average of the standard errors per sim, 'SD' is the sampling standard deviation of the mean. For each method in each draw, we calculated 95% normal intervals, and 'Coverage' reports the proportion of those which contained the true value, in this case calibrated to 3.

Figure 2: BART-IV Monte Carlo parameter estimates with confidence intervals across samples



We see a lower coverage proportion in the BART-IV confidence intervals than in the TSLS confidence intervals. So, the credible intervals from BART-IV do not accurately represent frequentist confidence intervals. We suspect that this is because the inference is driven by the choice of prior–the low coverage ratio here suggests that the prior is far off from the true DGP. Based on the result of this simulation, we conclude that without task-specific modification of the prior, BART-IV is inappopriate for use on the crime data.

4 Empirical Results

Having tested how well BART-IV and RF-IV are calibrated, we now turn to running these methods on the real crime data. For each method, our outcome variables of interest are violent crime and property crime, the aggregate categories used by Levitt, as well as the specific crimes that come under each category. In particular, murder, rape, assault, and robbery count as violent crimes, while burglary, larceny, and auto theft count as property crimes.

The results are presented in Table 3, where we report point estimates as well as standard deviation estimates. Below, we see estimates and standard deviations obtained using each of the four methods for the effect of $\Delta \ln(P_{it})$ on $\Delta \ln(C_{it})$, for various crimes C_{it} (which correspond to the rows of the table), as specified in section 2. For OLS and TSLS, the standard deviations, in columns (2) and (4) respectively, are the usual standard error estimates. For BART-IV, our estimate in column (5) is the mean of 50 draws from the posterior of the estimated coefficient, and our measure of standard deviation in column (5) is the standard deviation of these draws. For RF-IV, our estimate in column (7) is the mean of the RF-IV estimates obtained from 50 bootstrapped samples of the real data, while the measure in column (8) is the standard deviation of these RF-IV estimates.

Table 3: Parameter estimates using crime data. We use $\widehat{\ }$ to refer to the estimate using a certain method, and Δ to indicate the standard error in the estimate.

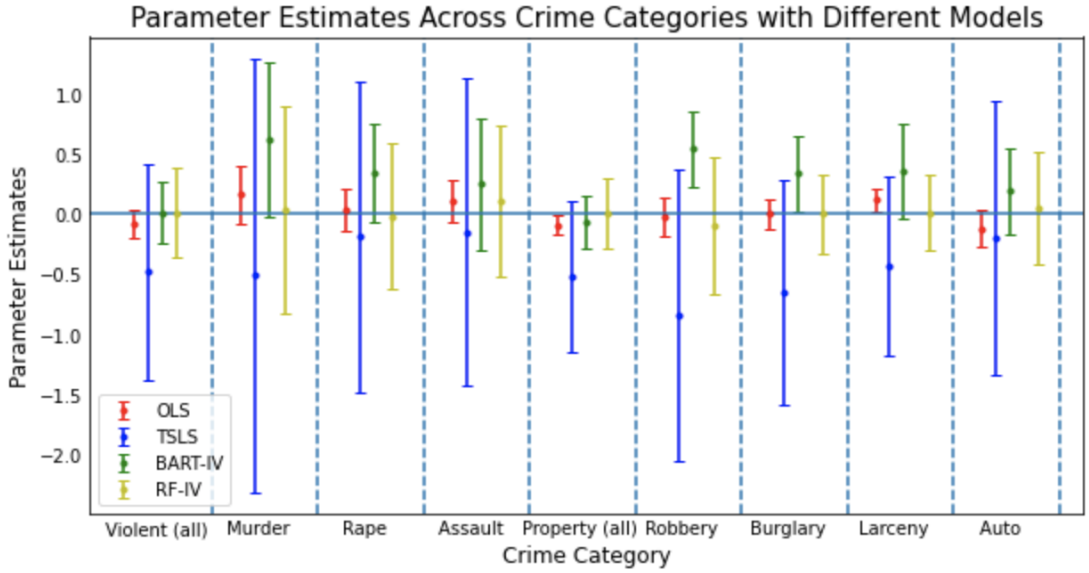
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Crime	$\widehat{\text{OLS}}$	ΔOLS	$\widehat{\text{TSLS}}$	ΔTSLS	$\widehat{\text{BART}}$	ΔBART	$\widehat{\text{RF}}$	ΔRF
Violent	-0.08	0.06	-0.48	0.46	0.01	0.13	0.01	0.19
Murder	0.16	0.123	-0.51	0.92	0.62	0.33	0.03	0.44
Rape	0.04	0.09	-0.19	0.66	0.34	0.21	-0.02	0.31
Assault	0.11	0.09	-0.15	0.65	0.25	0.28	0.11	0.32
Robbery	-0.02	0.08	-0.84	0.62	0.54	0.16	-0.09	0.29
Property	-0.09	0.04	-0.52	0.32	-0.07	0.11	0.01	0.15
Burglary	0.00	0.06	-0.65	0.48	0.34	0.16	0.00	0.17
Larceny	0.12	0.05	-0.43	0.38	0.36	0.20	0.01	0.16
Auto	-0.12	0.08	-0.20	0.58	0.19	0.18	0.05	0.24

We include the following controls for each method (W_{it} in section 2): Education spending per capita, public welfare spending per capita (both log-first-differenced), unemployment rate, proportion of population which is African-American, proportion of female headed-households, proportion of population between the ages of 15 and 24 (all first-differenced), year fixed effects (all methods), city fixed effects (all methods but RF).

In Figure 3, we depict estimates, as well as estimated 95% normal confidence intervals, $(\widehat{\text{est}} - 1.96 \cdot \Delta\text{est}, \widehat{\text{est}} + 1.96 \cdot \Delta\text{est})$. We see in Figure 3 that all of our estimated coefficients are indistinct from zero at a 95% confidence level (barring the

BART-IV estimate for robbery). All of our TSLS estimates (blue) have large standard errors, while both RF-IV and BART-IV have smaller standard errors. Based on our simulations, we expect RF-IV confidence intervals to be well-calibrated, and cover the true parameter with a success rate close to that of the TSLS. Our main finding in this empirical application, therefore, is that using a random forest to predict the first stage of Levitt's IV design gives us tighter confidence intervals around zero. Our RF-IV results agree with McCray's finding that the causal effect of police on the crime categories considered by Levitt is not statistically significant from zero, but we do so with significantly lower standard errors.

Figure 3: Visualization of the information in Table 3, using a 95% normal confidence interval



For the sake of completeness, we also include our BART-IV estimates and standard deviations. We do not claim to have used BART-IV successfully on Levitt's data, since we saw in section 3 that the Bayesian 95% credible intervals do not always correspond to the same confidence level in a frequentist setting, so we are unsure as to whether the BART-IV standard errors are well-calibrated. However, BART-IV demonstrates some potential to reduce TSLS standard errors – as seen in the simulations as well as the real data – which suggests directions for future research as discussed in section 5.

5 Discussion

Using the simulated data on our DGP, we saw that OLS had the tightest standard errors. However, because of the simultaneity problem, OLS does not yield a causal interpretation of the fitted parameter. By contrast, an instrument satisfying some conditions can yield a causal interpretation using the standard TSLS. However, the standard errors on this fitted parameter are high. As such, we explored different models, in particular the BART-IV and the RF-IV. We found that both typically reduced the standard error of the fitted parameter.

We found that the bootstrapped standard errors from RF-IV are well calibrated, but the Bayesian estimates of the BART-IV standard errors are uncalibrated because of sensitivity to prior choice. The sensitivity with respect to prior becomes less significant with a larger sample size, but it is something we need to be cognizant about when working with smaller datasets. It would be interesting to further investigate how to calibrate the Bayesian results from BART in smaller samples to obtain reliable inferences.

We also propose applications of RF-IV past crime and policing. As we observed, the RF-IV and the TSLS provided similar coverage in the simulation DGP, but with a more complicated true DGP, RF-IV clearly produced tighter confidence intervals. As such, RF-IV can be used in places where standard TSLS has been used to provide more precise confidence intervals, especially in the presence of more complicated DGPs or weaker instruments.

References

Angrist, J. and Frandsen, B. (2019). Machine labor. Technical report, National Bureau of Economic Research.

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87(418):328–336.

Ash, E., Chen, D., Zhang, X., Huang, Z., and Wang, R. (2018). Deep iv in law:

Analysis of appellate impacts on sentencing using high-dimensional instrumental variables.

Bai, J. and Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, pages 1577–1606.

Chen, J., Chen, D. L., and Lewis, G. (2020). Mostly harmless machine learning: Learning optimal instruments in linear iv models. *arXiv preprint arXiv:2011.06158*.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Levitt, S. D. (1995). Using electoral cycles in police hiring to estimate the effect of policeon crime. Technical report, National Bureau of Economic Research.

McCrary, J. (2002). Using electoral cycles in police hiring to estimate the effect of police on crime: Comment. *American Economic Review*, 92(4):1236–1243.

McCulloch, R. E., Sparapani, R. A., Logan, B. R., and Laud, P. W. (2021). Causal inference with the instrumental variable approach and bayesian nonparametric machine learning. *arXiv preprint arXiv:2102.01199*.